The Present Position of the Coding Problem

F.H(C.)Crick
MRC Unit for Molecular Biology, Cavendish Laboratory, Cambridge, England

It is widely assumed that the amino acid sequence of a particular protein is in some way determined by the sequence of the bases in some particular length of nucleic acid. While the *indirect* evidence in favor of some relationship of this type is very suggestive, the direct evidence is fragmentary in the extreme, and nothing whatever is known about the actual mechanisms involved. It is possible, however, to consider the problem in an abstract way as that of translating from one language to another; that is, from the 4-letter language of the nucleic acids to the 20-letter language of the protein, without any detailed consideration of the chemical processes involved. This approach is often referred to as the coding problem.

The coding problem has so far passed through three phases. In the first, the vague phase, various suggestions were made, but none was sufficiently precise to admit disproof. The second phase, the optimistic phase, was initiated by Gamow[1,2] in 1954, who was rash enough to suggest a fairly precise code. This stimulated a number of workers to show that his suggestion must be incorrect, and in doing so increased somewhat the precision of thinking in this field. The third phase, the confused phase, was initiated by the paper of Belozersky and Spirin[3] in 1958, although the experimental data had actually been published earlier, both by them[4] and by Lee, Wahl, and Barbu.[5] The evidence presented there showed that our ideas were in some important respects too simple. These difficulties will be the main topic of this paper.

The earlier work will not be reviewed here. That up to 1955 has been discussed by Gamow, Rich, and Ycas,[6] and some of the work since then has been briefly covered by Crick.[7] It suffices to say that Brenner[8] has shown by a study of amino acid sequences that all overlapping triplet codes, degenerate or not, are unlikely. (A code is called a triplet code if an amino acid is coded by a set of three consecutive bases; it is said to be an overlapping triplet code if any one base in the sequence forms part of the representations of three adjacent amino acids; and it is degenerate if some of the amino acids have more than one representation each.)

Several papers[9-12] have discussed the mathematics of commaless codes. The difficulties of constructing such codes applicable to the DNA double helix have been discussed by Golomb, Welch, and Delbruck,[11] who have been led to consider quadruplet codes. Brenner and Crick (unpublished) have done work along similar lines. However, until the present difficulties are overcome, most of these detailed schemes are not very pertinent. These difficulties first became widely appreciated upon the publication of the paper of Belozersky and Spirin.[3] They showed that the DNA of different microorganisms had greatly different base ratios. With the initial letters used to represent guanine, cytosine, adenine, thymine, and uracil, their results can be described by saying that the ratio (G+C)/(A+T) was as low as ½ for some

organisms and higher than 2½ for others. The base composition of the total RNA of these same organisms hardly varied at all, though there appeared to be a weak correlation between RNA composition and DNA composition. Similar results had been reported by Lee, Wahl, and Barbu.[5]

This large variation of DNA composition is very unexpected. The abundance of the various amino acids does not, as far as we know, vary much from organism to organism; leucine is always common, methionine usually rather rare. The small variation of RNA composition is exactly what might be expected; the large variation reported for DNA needs some special explanation. More recently work by Doty and his colleagues[13] and by Meselson (discussed elsewhere in this symposium) has established that for any one microorganism the base compositions of its different DNA molecules are all very similiar.

Listed below are some possible explanations of this phenomenon, though in my view they all, at the moment, appear unattractive. Some of these have been listed by Sueoka, Marmur, and Doty.[13]

1. Only Part of the DNA Codes Protein. It is postulated that the sequences of bases in a DNA molecule are of two types: one makes "sense," that is, codes an amino acid sequence; the other makes "nonsense," that is, has some other function. The difficulty of this idea is that the nonsense must make up a rather large fraction of the DNA. If, for example, it is assumed that the base composition of the sense is reflected in that of the total RNA of the organism, then organisms showing extreme base ratios must have a minimum of 35% nonsense in their DNA.

If nonsense exists it can be asked how, in one molecule of DNA, the sense and nonsense are interdispersed. Are they coarsely or finely dispersed? As an example of the former, consider what might happen if dud genes could not be eliminated by genetic deletion. The base composition of such genes might well drift to extreme values because of mutagenic bias within the cell. This explanation is not very likely, and in addition demands that dud genes be reasonably uniformly distributed among DNA molecules.

A possible reason for the fine dispersion of nonsense might be the provision of "commas." For example, these might take the form of segments that could pair by twisting back on themselves when the two chains of the DNA were separated. The base pairs of these regions could vary without altering their function. Alternatively a short sequence of bases, different from species to species but always the same in any one species, might act as a comma.

2. The DNA-to-RNA Translation Mechanism Varies. This would allow the RNA-to-protein code to be uniform throughout nature, while permitting the DNA-to-RNA code to vary. This is a formal possibility, but it does not seem at all likely. It has not been proposed in detail in any convincing form, and it is difficult to see how such a mechanism could be varied.

The remaining explanations all give reasons why the DNA might vary, but they might be expected to lead to parallel variations in the RNA. The fact that this is not observed has to be explained by some further hypothesis; for example, that only a fraction of the RNA of the cell is determined by the DNA, the remainder being produced in some other way, as suggested by Belozersky and Spirin[3] from the correlation shown by their experimental data.

3. The Code is Degenerate. This means that most amino acids have several representations. If these were very different in their base composition, it might be possible to account for the observed range of base ratios of the DNA.

4. The Code Is Not Universal. It used to be argued that the code would be uniform throughout nature (except possibly for certain virulent viruses) because any attempt to change it would necessarily alter many proteins at once and would thus almost certainly be lethal. However, a counterargument has been given (Levinthal, personal communication) that an alteration to a code need only be an extremely rare event, and that perhaps under certain conditions (for example, when the organism was in a rich environment and thus did not require too many enzymes) it might be possible to make a change. Most schemes of this type would also permit a change in one of the amino acids, and this appears not to have been observed. It does not seem that this point can usefully be argued. Some direct experimental evidence, perhaps from the soluble RNA, would be an advantage.

5. The Nucleic Acid Code Has Less Than Four Letters. In particular the code might be binary. Only three binary codes are possible. If adenine is equivalent to thymine, nothing is gained. It might be equivalent to guanine (in which case the two letters would be purine and pyrimidine), or it might be equivalent to cytosine (making the two letters 6-amino and 6-keto). This latter alternative is being proposed by Sinsheimer (in press). It has the advantage that, if RNA is made in the groove of the unaltered double helix of DNA, this degeneracy is structurally rather plausible, since it has proved impossible to devise schemes, from a study of models, which avoid it in any convincing way.

It is also possible to consider tertiary codes (i.e., having three letters), of which six types are possible, four of which would help overcome the difficulty. Without some other argument in their favor it cannot be said that these codes are very attractive.

6. The Amino Acid Composition of the Protein Varies. Unfortunately a small variation will not do. The organisms with extreme base ratios in their DNA are required to have proteins for which, say, leucine is rare and methionine common. This possibility should be tested experimentally, but it does not seem very likely.

It remains to consider briefly by what means the problem might be attacked experimentally. The obvious long-term approaches are either by way of the soluble RNA, as discussed by Hoagland and by Brown in this symposium, or by the study of the gene-protein interrelationship, discussed here by Levinthal and by Brenner, combined with specific mutagenesis, of which Freese has given an account.

A useful short-term approach would be to study the special RNA fractions from organisms with extreme base ratios in their DNA. At least three significant fractions are known: the soluble RNA, the RNA in the larger ribosomal component, and the RNA in the smaller ribosomal component. It would be of considerable interest if it were found that the base ratios of one of these fractions followed those of the DNA. It would also be of interest to know whether the terminal sequence of the soluble RNA is always ACC.

Obviously it would be an advantage to have some sequence information for the DNA molecules of extreme base composition, and especially any evidence of re-

peating sequences. Sonication of the DNA molecules (as briefly reported[13]) might show whether smaller lengths of DNA have a fairly uniform base composition.

Finally, lest the reader be too discouraged, a few important experimental facts should be mentioned which any theory will have to explain: first, the evidence that the RNA of tobacco mosaic virus controls, at least in part, the amino acid sequence of the protein of the virus; second, the genetic effects of transforming factor, which appears to be pure DNA; and third, the genetic control of at least parts of the amino acid sequence of human hemoglobin.

REFERENCES

1. Gamow, G., *Nature* **173**, 318 (1954).
2. Gamow, G., *Biol. Medd. Dan. Vid. Selsk.* **22**, 3 (1954).
3. Belozersky, A.N. and Spirin, A.S., *Nature* **182**, 111 (1958).
4. Spirin, A.S., Belozersky, A.N., Shugaeva, N.V., and Vanjushin, B.F., *Biokhimiya* **22**, 744 (1957).
5. Lee, K.Y., Wahl, R., and Barbu, E., *Ann. inst. Pasteur* **91**, 212 (1956).
6. Gamow, G., Rich, A., and Ycas, M., in *Advances in Biological and Medical Physics*, IV, J.H. Lawrence and J.G. Lawrence, Editors, Academic Press, New York, 1955.
7. Crick, F.H.C., in *The Biological Replication of Macromolecules, Symp. Soc. Exptl. Biol.* **12**, 138 (1958).
8. Brenner, S., *Proc. Natl. Acad. Sci. U.S.* **43**, 687 (1957).
9. Crick, F.H.C., Griffith, J.S., and Orgel, L.E., *Proc. Natl. Acad. Sci. U.S.* **43**, 416 (1957).
10. Golomb, S.W., Gordon, B., and Welch, L.R., *Can. J. Math.* **10**, 202 (1958).
11. Golomb, S.W., Welch, L.R., and Delbruck, M., *Biol. Medd. Dan. Vid. Selsk.* **23**, 1 (1958).
12. Freudenthal, H., *Koninkl. Ned. Akad. Wetenschap. Proc.* **A61**, 253 (1958).
13. Sueoka, N., Marmur, J., and Doty, P., *Nature* **183**, 1429 (1959).

DISCUSSION

Zamenhof: I would like to comment on the fact that we don't have four bases in either RNA or DNA. In RNA the latest count is eight, and it goes up all the time. Would you have too low a concentration of information if it turns out that you have almost as many bases as amino acids? Lately it was also found that there are some differences in sugar. Could it be that all these things are not gene determined? Or could it be that there is some information also in the sugar, God forbid?

Crick: Hinshelwood once suggested a code dependent on the sugar but I never understood what he meant. We would be much happier, of course, if there were a third form of base-pairing, because of the problem of specific replication. If the replication of DNA goes the way we expect, with just the two base pairs, then the extra bases may not be adding information. The occurrence of these bases does not seem to be correlated with the amino acid composition of the proteins of the cell in a systematic way. The most graphic example is that of the 5-hydroxymethylcytosine of the T phages, with the glucose on it. This is a big change, and yet the proteins seem to be the same kind of proteins. Moreover, we know that Kornberg's system handles these precursors in the way one would expect. In the case of 5-methylcytosine the difficulties are greater, because the 5-methylcytosine might be expected to go in at random, but it doesn't. If we accept, tentatively, Sinsheimer's evidence from the enzymatic digestion, it would look as if 5-methylcytosine always occurs next to guanine. This does not necessarily give more information in the technical sense. In general we can see no correlation between unusual bases and amino acid composition.

I should make two further points. Unusual bases are of three types. In one type a methyl group is added, and the base pairing is unaffected. There is no case of "good" nucleic acid, by which I mean DNA or virus RNA, having a base which does not pair. There is a second class

of base, which won't base pair, but these are not found in the RNA of tobacco mosaic virus or in DNA. There is a third class, which consists of pseudo-uracil, and that certainly is not found in TMV RNA but it is found in soluble RNA; I don't want to go into that now because I presume Dr. Hoagland will discuss it. The real reason we ignore these extra bases is that no one can fit them into any meaningful pattern. If you can find a way of doing so, that would be fine.

Muller: I believe it is implied in Chargaff's rule that the total adenine plus uracil of RNA maintains a constant ratio to the total guanine plus cytosine. Couldn't this be most simply interpreted on the supposition that the RNA has two kinds of strands like the DNA but that the proportions of these strands may vary, that one may become duplicated much more than the other? If that were true, then, where there was a given pyrimidine in one, there would be a complementary purine in the other, so that any variation in relative numbers of the complementary strands would keep the total amount of adenine plus uracil the same in relation to the total amount of guanine plus cytosine.

Crick: Your suggestion is that the RNA is made by base pairing but that single chains are thrown off. This might be possible. An alternative explanation is that in RNA synthesis the DNA type of pairing occurs but that there is no restriction on the distance apart of the backbones. Rich already has a pair A+I which is not very different from A+G. If you allow the usual pairs, that is, adenine plus uracil and guanine plus cytosine, and in addition A+G and C+U, then any structure that has those four pairings will give you Chargaff's rule, which is that 6-keto equals 6-amino. We suspect that in DNA replication the enzyme may put this restriction on the distance apart of the chains and thus restrict the pairing. In RNA synthesis it might not do this.

Sager: I would like to suggest another way of looking at this greater variability in the DNA than in the RNA. This is based on a model that is relatively unpopular at present, but this may be all to the good. This is a model that proposes that a significant part of the information for coding protein is carried in RNA autonomously. In other words, there is a big fraction that is genetic RNA, and the relatively greater similarity that would be detected between over-all RNA and over-all protein is a reflection of this, and DNA contributes only a small but critical part of the total information, and therefore there would be much greater possibility for variability in that fraction.

Crick: Well that is a perfectly reasonable model. One might give a particular example of it by postulating that what the RNA is mainly coding is the actual protein of the ribosomes, but there does not seem to be base pairing in this RNA, which is embarrassing.

Fraser: What about an insane protein like poly-D-glutamic acid?

Crick: Sir, poly-D-glutamic acid is not a protein!